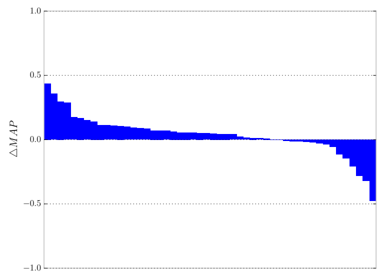
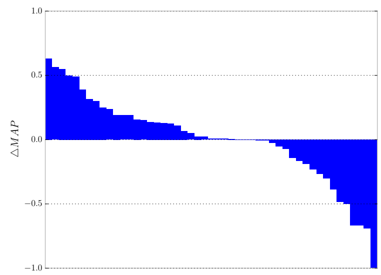
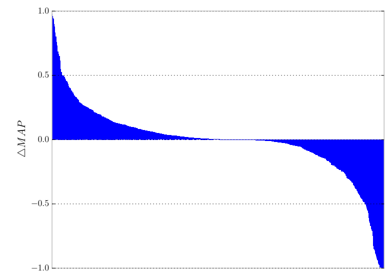


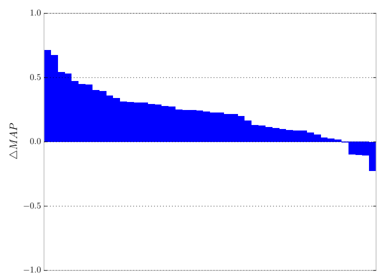
(a) W3C 2005



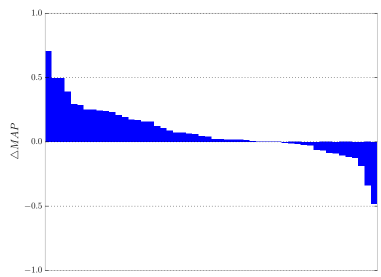
(b) CERC 2007



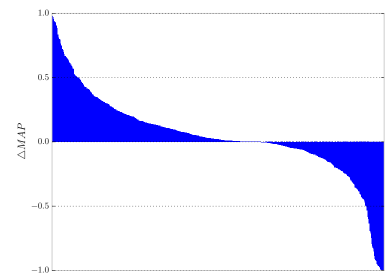
(c) TU GT1



(d) W3C 2006



(e) CERC 2008



(f) TU GT5

Figure 3: Difference of average precision between log-linear model and Model 2 [4] with Jelinek-Mercer smoothing per topic for W3C, CERC and TU benchmarks.

picked at random such that the batched update rule (4) behaves like the empirical expectation over the full training set [11]. While we might be able to justify the assumption that documents arrive randomly, the n-grams extracted from those documents clearly violate this requirement.

Considering a stream of documents leads to the model forgetting expertise evidence as an (artificial) shift in the underlying distribution of the training data occurs. While such behavior is undesirable for the task considered in this paper, it might be well-suited for temporal expert finding [22, 52], where expertise drift over time is considered. However, temporal expertise finding is beyond the scope for this paper and left for future work.

6. CONCLUSIONS

We have introduced an unsupervised discriminative, log-linear model for the expert retrieval task. Our approach exclusively employs raw textual evidence. Future work can focus on improving performance by feature engineering and incorporation of external evidence. Furthermore, no relevance feedback is required during training. This renders the model suitable for a broad range of applications and domains.

We evaluated our model on the W3C, CERC and TU benchmarks and compared it to state-of-the-art vector space-based entity ranking (based on LSI and TF-IDF) and language modeling (profile-centric and document-centric) approaches. The log-linear model combines the ranking performance of the best maximum-likelihood language modeling approach (document-centric) with inference time complexity linear in the number of candidate experts. We observed a notable increase in precision over existing methods. Analysis of our model's output reveals a negative correlation between the per-query performance and ranking uncertainty: higher confidence (i.e., lower entropy) in the rankings produced by our approach often occurs together with higher rank quality.

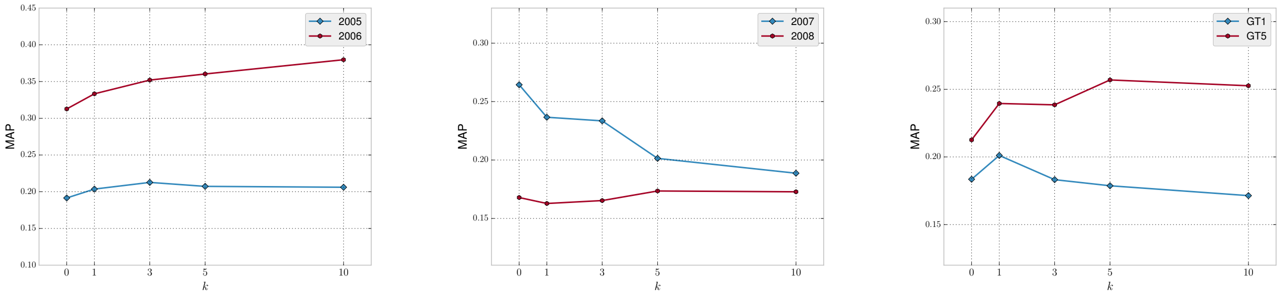
An error analysis of the log-linear model and traditional language models shows that the two make very different errors. These errors are mainly due to the semantic gap between query intent and the raw textual evidence. Some benchmarks expect exact query matches, others are helped by our semantic matching. An ensemble of methods employing exact and semantic matching generally outperforms the individual methods. This observation calls for further research in the area of combining exact and semantic matching.

One current limitation of our work is its scalability with respect to the number of candidate experts. We have started investigating trade-offs between model performance and time/space complexity. In the future we hope to apply scalable variants of this method on large-scale social media communities, for the purpose of determining topic ownership. While in this work we focus on expertise retrieval, the ideas we proposed can easily be transferred to the more general entity retrieval task. Moreover, our approach is likely to be applicable to authorship attribution and various other entity retrieval and prediction tasks.

Acknowledgments. We thank Isaac Sijaranamual, Manos Tsagkias, Tom Kenter, Zhaochun Ren and Ke Tran for their useful comments and insights.

This research was supported by Amsterdam Data Science, the Dutch national program COMMIT, Elsevier, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the ESF Research Network Program ELIAS, the Royal Dutch Academy of Sciences (KNAW) under the Elite Network Shifts project, the Microsoft Research Ph.D. program, the Netherlands eScience Center under project number 027.012.105, the Netherlands Institute for Sound and Vision, the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, 612.066.-930, CI-14-25, SH-322-15, 652.002.001, 612.001.551, the Yahoo Faculty Research and Engagement Program, and Yandex. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Computing resources were provided by the Netherlands Organisation for Scientific Research (NWO) through allocation SH-322-15 of the Cartesius system and by the Advanced School for Computing and Imaging (ASCII) by allocation of the Distributed ASCII Supercomputer 4 (DAS-4) system.



(a) W3C **(b) CERC** **(c) TU**

Figure 4: Effect of query expansion by adding nearby terms in W_p (1) in traditional language models (Model 1 [4] with Jelinek-Mercer smoothing) for W3C, CERC and TU benchmarks.

APPENDIX

The derivative of (3) w.r.t. bias term $\boldsymbol{b_c}$ equals

$$\frac{\partial L(W_p, W_c, \boldsymbol{b_c})}{\partial \boldsymbol{b_c}} = -\frac{1}{m}\left(\sum_{i=1}^{m} \frac{|d_{\max}|}{|d^{(i)}|} \sum_{j=1}^{|C|} P(c_j|d^{(i)}) \frac{\partial \log\left(P(c_j \mid w_1^{(i)}, \ldots, w_n^{(i)})\right)}{\partial \boldsymbol{b_c}}\right)$$

and w.r.t. an arbitrary matrix parameter θ (W_p or W_c):

$$\begin{aligned}\frac{\partial L(W_p, W_c, \boldsymbol{b_c})}{\partial \theta} &= -\frac{1}{m}\left(\sum_{i=1}^{m} \frac{|d_{\max}|}{|d^{(i)}|} \sum_{j=1}^{|C|} P(c_j|d^{(i)}) \frac{\partial \log\left(P(c_j \mid w_1^{(i)}, \ldots, w_n^{(i)})\right)}{\partial \theta}\right) \\ &\quad + \frac{\lambda}{m}\sum_{i,j}\theta_{i,j}.\end{aligned}$$

Further differentiation for parameter θ (W_p, W_c or $\boldsymbol{b_c}$):

$$\begin{aligned}\frac{\partial \log\left(P(c_j \mid w_1, \ldots, w_n)\right)}{\partial \theta} &= \frac{1}{P(c_j \mid w_1, \ldots, w_n)} \frac{\partial P(c_j \mid w_1, \ldots, w_n)}{\partial \theta} \\ \frac{\partial P(c_j \mid w_1, \ldots, w_n)}{\partial \theta} & \\ &= \frac{\frac{\partial \tilde{P}(c_j|w_1,\ldots,w_n)}{\partial \theta} Z_2 - \tilde{P}(c_j \mid w_1, \ldots, w_n)\frac{\partial Z_2}{\partial \theta}}{Z_2^2} \\ \frac{\partial Z_2}{\partial \theta} &= \sum_k \frac{\partial \tilde{P}(c_k \mid w_1, \ldots, w_n)}{\partial \theta} \\ \frac{\partial \tilde{P}(c_j \mid w_1, \ldots, w_n)}{\partial \theta} &= \sum_k \frac{\partial P(c_j \mid w_k)}{\partial \theta} \prod_{i \neq k} P(c_j \mid w_i)\end{aligned}$$

For a given candidate c_j and word w_i, following (1) we have

$$\begin{aligned}P(c_j \mid w_i) &= \frac{\tilde{P}(c_j \mid w_i)}{Z_1} \\ &= \frac{\exp\left((\sum_{k=1}^{e} W_{c_{j,k}} W_{p_{k,i}}) + b_{c_j}\right)}{\sum_{l=1}^{|C|} \exp\left((\sum_{k=1}^{e} W_{c_{l,k}} W_{p_{k,i}}) + b_{c_l}\right)}\end{aligned}$$

and consequently, with $\boldsymbol{W_{p_i}^{\top}}$ denoting the i-th column of matrix W_p,

$$\begin{aligned}\frac{\partial P(c_j \mid w_i)}{\partial \boldsymbol{W_{c_j}}} &= \frac{\left(Z_1 - \tilde{P}(c_j \mid w_i)\right)\tilde{P}(c_j \mid w_i)\boldsymbol{W_{p_i}^{\top}}}{Z_1^2} \\ \frac{\partial P(c_j \mid w_i)}{\partial b_{c_j}} &= \frac{\left(Z_1 - \tilde{P}(c_j \mid w_i)\right)\tilde{P}(c_j \mid w_i)}{Z_1^2} \\ \frac{\partial P(c_j \mid w_i)}{\partial \boldsymbol{W_{p_i}^{\top}}} &= \frac{\left(\boldsymbol{W_{c_j}} - \sum_{l=1}^{|C|}\boldsymbol{W_{c_l}}\right)\tilde{P}(c_j \mid w_i)}{Z_1}\end{aligned} \qquad (7)$$

As can be seen in (7), the distributed representations of candidates c_j at time $t+1$ are updated using the representation of words w_i at time t and vice versa.

REFERENCES

[1] The knowledge-based economy. Techn. report, Organisation for Economic Co-operation and Development, 1996.

[2] P. Bailey, A. P. De Vries, N. Craswell, and I. Soboroff. Overview of the TREC 2007 enterprise track. In *TREC*, 2007.

[3] K. Balog. *People Search in the Enterprise*. PhD thesis, University of Amsterdam, 2008.

[4] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR*, pages 43–50, 2006.

[5] K. Balog, L. Azzopardi, and M. de Rijke. A language modeling framework for expert finding. *IPM*, 45:1–19, 2009.

[6] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. Expertise retrieval. *Found. & Tr. in Information Retrieval*, 6 (2-3):127–256, 2012.

[7] I. Becerra-Fernandez. Role of artificial intelligence technologies in the implementation of People-Finder knowledge management systems. *Knowledge-Based Systems*, 13(5):315–320, 2000.

[8] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *JMLR*, 3:1137–1155, 2003.

[9] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JSTOR*, pages 289–300, 1995.

[10] R. Berendsen, M. de Rijke, K. Balog, T. Bogers, and A. van den Bosch. On the assessment of expertise profiles. *JASIST*, 64(10): 2024–2044, 2013.

[11] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, pages 177–186. Springer, 2010.

[12] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, 2005.

[13] Y. Cao, J. Liu, S. Bao, and H. Li. Research on Expert Search at Enterprise Track of TREC 2005. In *TREC*, pages 2–5, 2005.

[14] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *JMLR*, 12(Aug):2493–2537, 2011.

[15] N. Craswell, D. Hawking, A.-M. Vercoustre, and P. Wilkins. P@noptic expert: Searching for experts not just for documents. In *Ausweb Poster Proceedings*, pages 21–25, 2001.

[16] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC 2005 enterprise track. In *TREC*, 2005.

[17] T. H. Davenport and L. Prusak. *Working Knowledge*. Harvard Business Review Press, 1998.

[18] S. C. Deerwester, S. T. Dumais, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 1990.

[19] G. Demartini, J. Gaugaz, and W. Nejdl. A vector space model for ranking entities and its application to expert search. In *ECIR*, pages 189–201. Springer, 2009.

[20] L. Deng, X. He, and J. Gao. Deep stacking networks for information retrieval. In *ICASSP*, pages 3153–3157, 2013.

[21] H. Fang and C. Zhai. Probabilistic models for expert finding. In *ECIR*, pages 418–430, Berlin, Heidelberg, 2007. Springer-Verlag.

[22] Y. Fang and A. Godavarthy. Modeling the dynamics of personal expertise. In *SIGIR*, pages 1107–1110, 2014.

[23] Y. Fang, L. Si, and A. P. Mathur. Discriminative models of integrating document evidence and document-candidate associations for expert search. In *SIGIR*, pages 683–690, 2010.

[24] K. Fatahalian, J. Sugerman, and P. Hanrahan. Understanding the efficiency of gpu algorithms for matrix-matrix multiplication. In *SIGGRAPH HWWS*, pages 133–137. ACM, 2004.

[25] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, pages 249–256, 2010.

[26] G. E. Hinton. Learning distributed representations of concepts. In *8th Annual Conference of the Cognitive Science Society*, volume 1, page 12, Amherst, MA, 1986.

[27] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57. ACM, 1999.

[28] P.-s. Huang, N. M. A. Urbana, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *CIKM*, pages 2333–2338, 2013.

[29] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *STOC*, pages 604–613. ACM, 1998.

[30] M. Karimzadehgan and C. Zhai. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In *SIGIR*, pages 323–330. ACM, 2010.

[31] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *ICML*, pages 595–603, 2014.

[32] J. Kruger and D. Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Personality and Social Psych.*, 77 (6):1121, 1999.

[33] J. Krüger and R. Westermann. Linear algebra operators for gpu implementation of numerical algorithms. *ACM Transactions on Graphics*, 22(3):908–916, 2003.

[34] H. Li and J. Xu. Semantic matching in search. *Found. & Tr. in Information Retrieval*, 7(5):343–469, June 2014.

[35] T.-Y. Liu. *Learning to Rank for Information Retrieval*. Springer, 2011.

[36] C. MacDonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM*, pages 387–396, 2006.

[37] C. Macdonald and I. Ounis. Expert search evaluation by supporting documents. In *ECIR*, pages 555–563. Springer, 2008.

[38] M. T. Maybury. Expert finding systems. Techn. Report MTR-06B000040, MITRE, 2006.

[39] D. W. McDonald and M. S. Ackerman. Expertise recommender. In *CSCW*, pages 231–240, 2000.

[40] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur. Recurrent neural network based language model. In *Interspeech*, pages 1045–1048, 2010.

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[42] T. Mikolov, G. Corrado, K. Chen, and J. Dean. Efficient estimation of word representations in vector space. arXiv 1301.3781, 2013.

[43] A. Mnih and G. Hinton. Three new graphical models for statistical language modelling. In *ICML*, pages 641–648, 2007.

[44] A. Mnih and G. Hinton. A scalable hierarchical distributed language model. In *NIPS*, pages 1081–1088, 2008.

[45] A. Mnih and K. Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *NIPS*, pages 2265–2273, 2013.

[46] G. Montavon, G. B. Orr, and K.-R. Müller. *Neural Networks: Tricks of the Trade*. Springer, 2012.

[47] C. Moreira, B. Martins, and P. Calado. Using rank aggregation for expert search in academic digital libraries. In *Simpósio de Informática, INForum*, pages 1–10, 2011.

[48] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global Vectors for Word Representation. In *EMNLP*, pages 1532–1543, 2014.

[49] D. Petkova and W. B. Croft. Hierarchical language models for expert finding in enterprise corpora. In *ICTAI '06*, pages 599–606, 2006.

[50] W. W. Powell and K. Snellman. The knowledge economy. *Annual review of sociology*, pages 199–220, 2004.

[51] D. Rumelhart, G. Hinton, and R. Williams. Learning internal representations by back propagation. In *Parallel Distributed Processing*, pages 318–362. MIT Press, 1986.

[52] J. Rybak, K. Balog, and K. Nørvåg. Temporal expertise profiling. In *ECIR*, pages 540–546. Springer, 2014.

[53] R. Salakhutdinov and G. Hinton. Semantic hashing. *Int. J. Approximate Reasoning*, 50(7):969–978, 2009.

[54] P. Serdyukov and D. Hiemstra. Modeling documents as mixtures of persons for expert finding. In *ECIR*, pages 309–320. Springer, 2008.

[55] P. Serdyukov, H. Rode, and D. Hiemstra. Modeling multi-step relevance propagation for expert finding. In *CIKM*, pages 1133–1142, 2008.

[56] C. Shannon. A mathematical theory of communication. *Bell System Technical J.*, 27:379–423, 623–656, 1948.

[57] J. A. Shaw, E. A. Fox, J. A. Shaw, and E. A. Fox. Combination of multiple searches. In *TREC*, pages 243–252, 1994.

[58] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. A latent semantic model with convolutional-pooling structure for information retrieval. In *CIKM*, pages 101–110, 2014.

[59] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM*, pages 623–632. ACM, 2007.

[60] P. Sorg and P. Cimiano. Finding the right expert: Discriminative models for expert retrieval. In *KDIR*, pages 190–199, 2011.

[61] TREC. Enterprise Track, 2005–2008.

[62] D. van Dijk, M. Tsagkias, and M. de Rijke. Early detection of topical expertise in community question and answering. In *SIGIR*, 2015.

[63] V. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[64] M. D. Zeiler. Adadelta: An adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.